KU LEUVEN

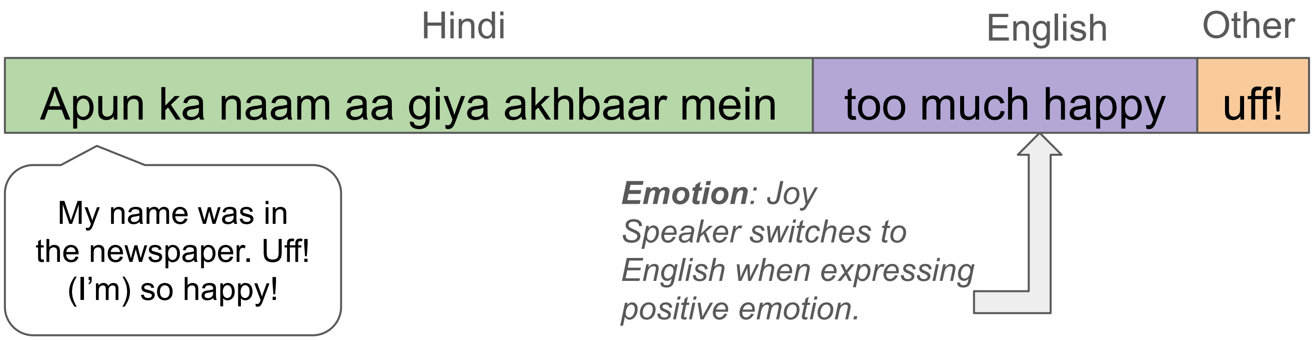
# Sociolinguistically Informed Interpretability:
## A Case Study on Hinglish Emotion Classification

**Kushal Tatariya**
Heather Lent
Johannes Bjerva
Miryam de Lhoneux

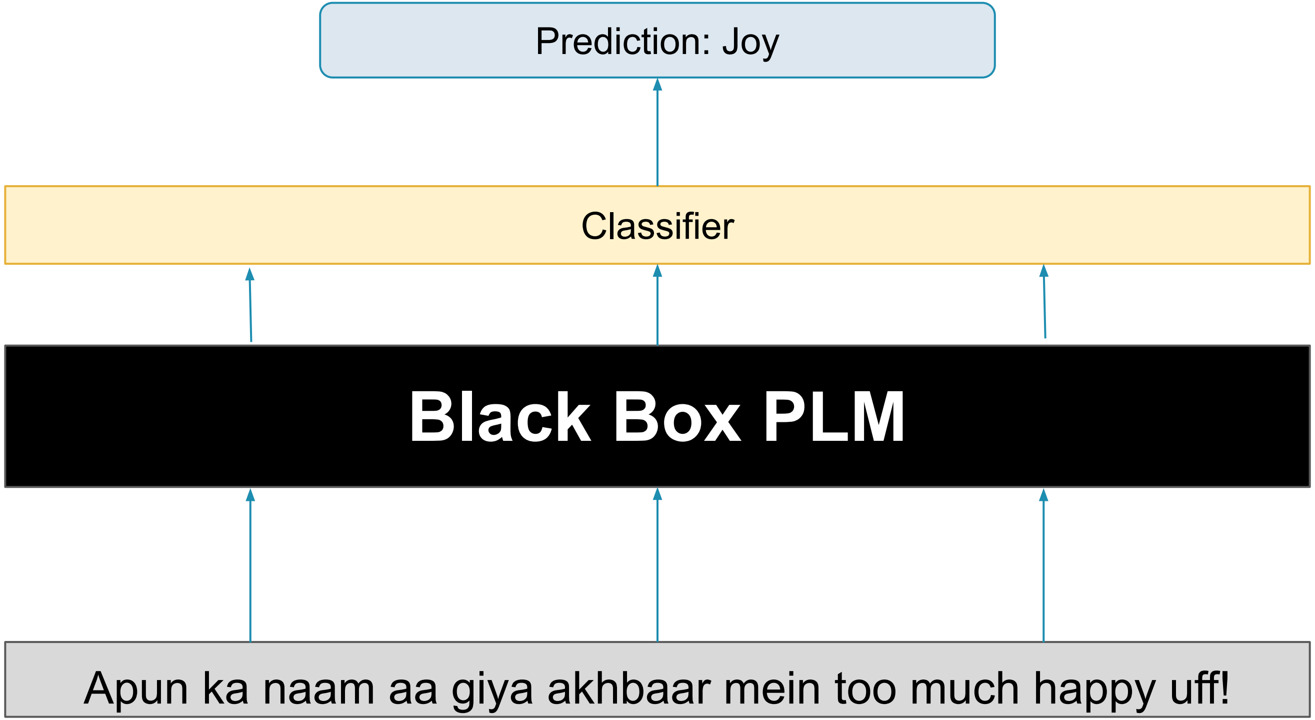
| | | |
|---|---|---|
| Apun ka naam aa giya akhbaar mein | too much happy | uff! |

Hindi | English | Other
---|---|---
Apun ka naam aa giya akhbaar mein | too much happy | uff!

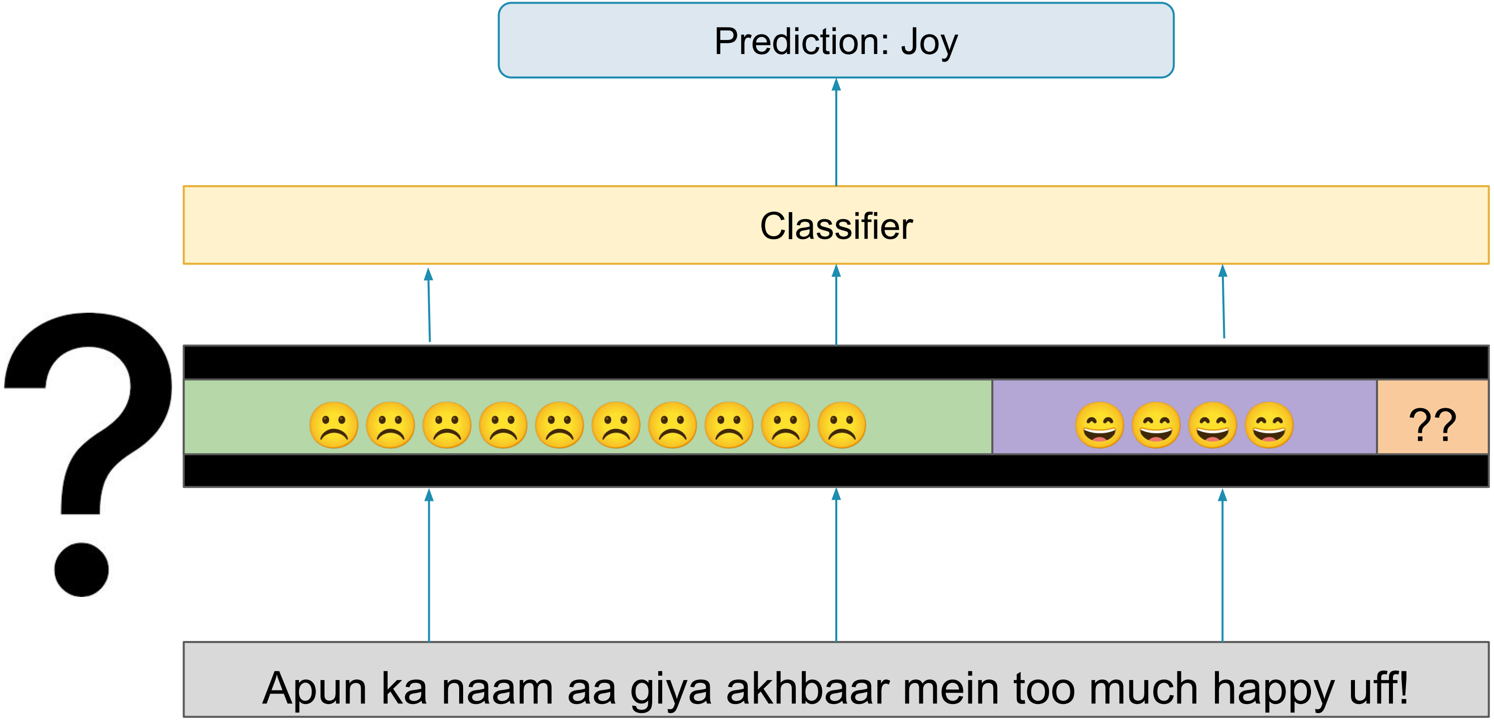
My name was in the newspaper. Uff! (I’m) so happy!


Hindi
English
Other
Apun ka naam aa giya akhbaar mein
too much happy
uff!
My name was in the newspaper. Uff! (I’m) so happy!
**Emotion**: Joy
Speaker switches to English when expressing positive emotion.

# Emotion classification pipeline



# Emotion classification pipeline






KU LEUVEN

# Concretely…

# Concretely…

**Question 1**

# Concretely…

**Question 1**

Are English tokens more likely to influence a model to predict a positive emotion?

# Concretely…

**Question 1**

Are English tokens more likely to influence a model to predict a positive emotion?

**Question 2**

# Concretely…

**Question 1**

Are English tokens more likely to influence a model to predict a positive emotion?

**Question 2**

Are Hindi tokens more likely to influence a model to predict a negative emotion? And if so, what is the role of Hindi swear words?

# Methodology

1. **Fine-tuning**

- XLM-R, IndicBERT and HingRoBERTa on a Hinglish emotion classification dataset (Ghosh et al, 2023).

# Methodology

1. **Fine-tuning**

- XLM-R, IndicBERT and HingRoBERTa on a Hinglish emotion classification dataset (Ghosh et al, 2023).
- 7 emotion labels: *Joy*, *Surprise*, *Sadness*, *Fear*, *Disgust*, *Anger*, *Other*

# Methodology

1. **Fine-tuning**

- XLM-R, IndicBERT and HingRoBERTa on a Hinglish emotion classification dataset (Ghosh et al, 2023).
- 7 emotion labels: *Joy*, *Surprise*, *Sadness*, *Fear*, *Disgust*, *Anger*, *Other*

| **Label distribution in the dataset** | | | | | | |
|---|---|---|---|---|---|---|
| Positive Emotions | | Neutral | Negative Emotions | | | |
| Joy | Surprise | Others | Anger | Sadness | Disgust | Fear |
| 33% | 2% | 35% | 20% | 10% | 19% | 1% |

# Methodology

## 1. Fine-tuning

- XLM-R, IndicBERT and HingRoBERTa on a Hinglish emotion classification dataset (Ghosh et al, 2023).
- 7 emotion labels: *Joy*, *Surprise*, *Sadness*, *Fear*, *Disgust*, *Anger*, *Other*

| Label distribution in the dataset | | | | | | |
|---|---|---|---|---|---|---|
| Positive Emotions | | Neutral | Negative Emotions | | | |
| Joy | Surprise | Others | Anger | Sadness | Disgust | Fear |
| 33% | 2% | 35% | 20% | 10% | 19% | 1% |

## 2. Token-Tagging

1000 samples, stratified across labels, each token annotated with:

## 2. Token-Tagging

1000 samples, stratified across labels, each token annotated with:

- Token level language ID : English, Hindi, Other

## 2. Token-Tagging

1000 samples, stratified across labels, each token annotated with:

- Token level language ID : English, Hindi, Other
- LIME score: Between -1 and 1

## 2. Token-Tagging

1000 samples, stratified across labels, each token annotated with:

- Token level language ID : English, Hindi, Other
- LIME score: Between -1 and 1
    - A positive score = token influenced the model *towards* the predicted label.

## 2. Token-Tagging

1000 samples, stratified across labels, each token annotated with:

- Token level language ID : English, Hindi, Other
- LIME score: Between -1 and 1
    - A positive score = token influenced the model *towards* the predicted label.
    - A negative score = token influenced the model to *not* predict that label.

Prediction: Joy

Hin Hin Hin Hin Hin Hin Hin Eng Eng Eng Other

Apun ka naam aa giya akhbaar mein too much happy uff!

-0.09 -0.08 0.02 -0.01 0.07 -0.02 -0.07 0.08 0.2 0.5 0.07

Prediction: Joy

Hin Hin Hin Hin Hin Hin Hin Eng Eng Eng Other

Apun ka naam aa giya akhbaar mein too much happy uff!

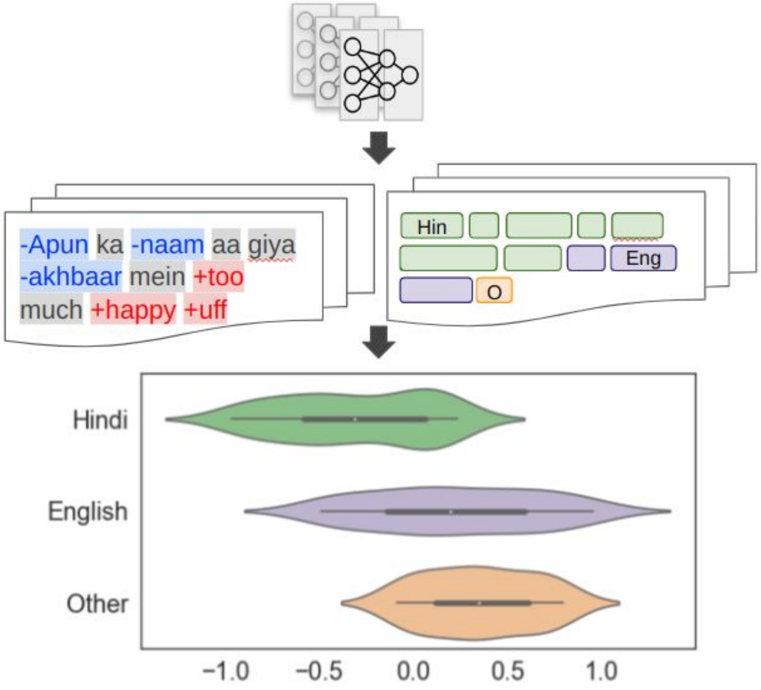
-0.09 -0.08 0.02 -0.01 0.07 -0.02 -0.07 0.08 0.2 0.5 0.07

Negative LIME score - influences model to not predict that label

Positive LIME score - influences model towards predicted label

### 3. Statistical Significance

- Frequency of positive/negative LIME score per language ID tag, for each model.
- Statistical testing with chi-square.



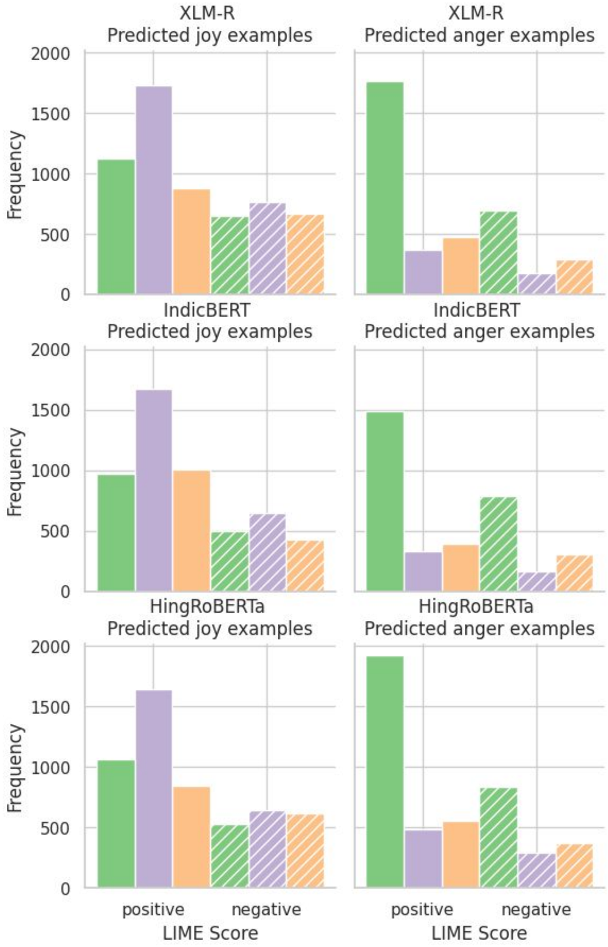
# Results

For predicted *joy* and *anger* examples:

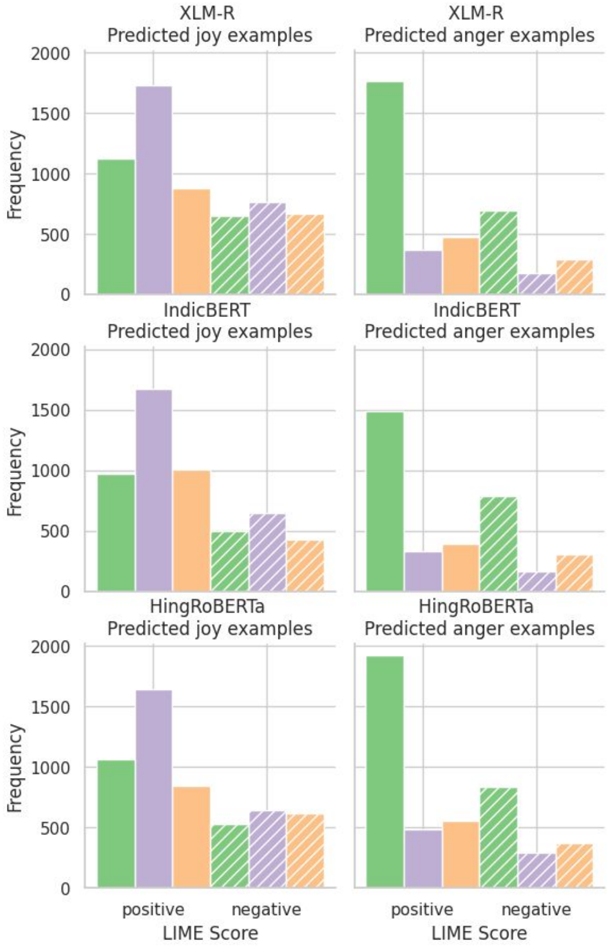
- All p-values are <0.05
- There is dependency between language ID and LIME score

| | $p$-values | | |
|---|---|---|---|
| **Model** | **Entire Sample** | **Joy** | **Anger** |
| **XLM-R** | 7.06e-12 | 1.44e-15 | 6.18e-7 |
| **IndicBERT** | 1.22e-22 | 3.28e-4 | 1.69e-5 |
| **HingRoBERTa** | 3.30e-7 | 4.00e-18 | 1.71e-8 |

We test the null hypothesis that language ID tags and LIME scores are independent of each other using χ2 .


XLM-R
Predicted joy examples
XLM-R
Predicted anger examples
IndicBERT
Predicted joy examples
IndicBERT
Predicted anger examples
HingRoBERTa
Predicted joy examples
HingRoBERTa
Predicted anger examples
Frequency
2000
1500
1000
500
0
positive
negative
LIME Score

# Do English tokens influence models to predict positive emotions?



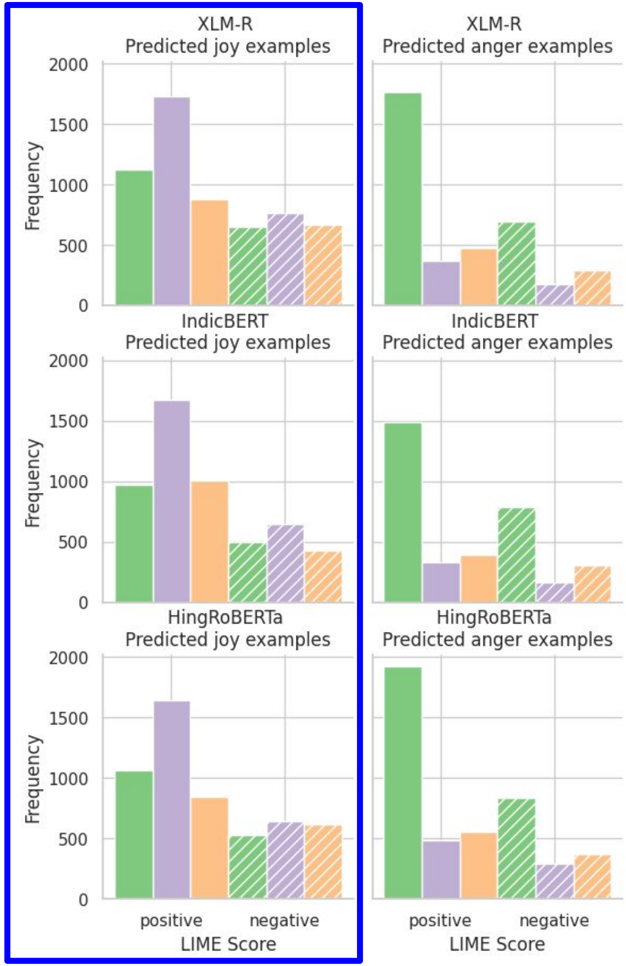



KU LEUVEN

# Do English tokens influence models to predict positive emotions?

## Yes!

English > Hindi, Other





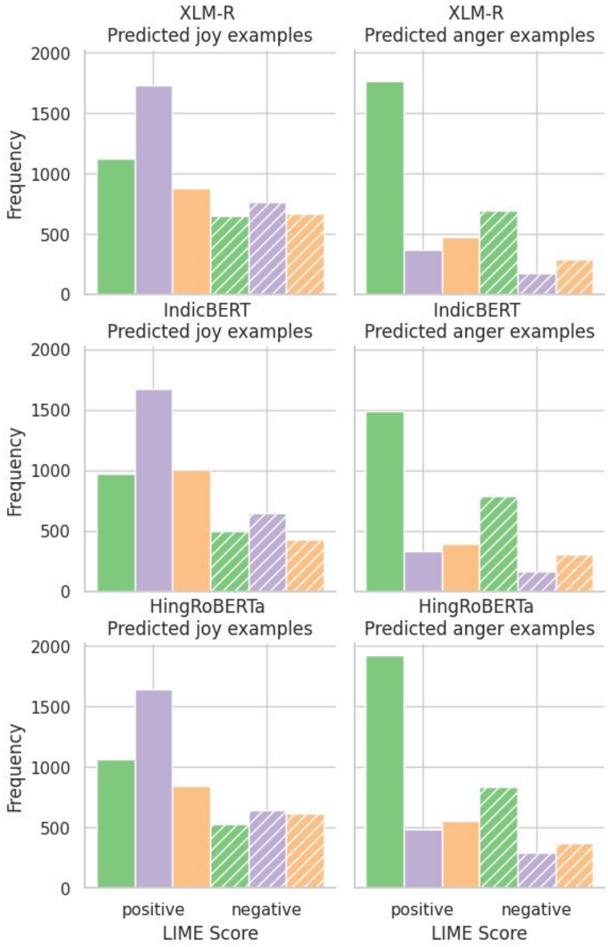

KU LEUVEN

## Do English tokens influence models to predict positive emotions?

**Yes!**

**English > Hindi, Other**

## Do Hindi tokens influence models to predict negative emotions?






KU LEUVEN

## Do English tokens influence models to predict positive emotions?

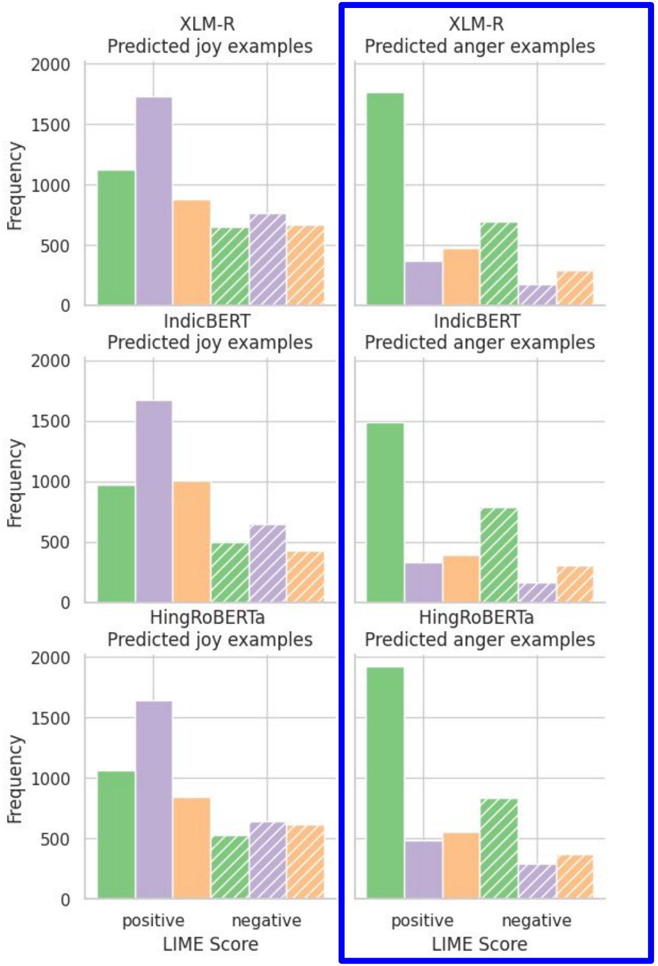
**Yes!**

**English > Hindi, Other**

## Do Hindi tokens influence models to predict negative emotions?

**Yes!**

**Hindi > English, Other**






KU LEUVEN

## What is the role of Hindi swear words?

| Token | Lang_ID | Swear Word?[2] |
|---|---|---|
| Fuck | eng | Yes |
| Chutiye | hin | Yes |
| Fakeionist | eng | No |
| Bsdk | hin | Yes |
| Sadly | eng | No |
| Bakwas | hin | No |
| Kutta | hin | Yes |
| Gaddar | hin | No |
| Shame | eng | No |
| Sala | hin | Yes |

Top 10 tokens with the highest LIME scores when predicting negative emotions, (anger, sadness, disgust and fear) for all models. They have been mapped to a canonical form and are in descending order of LIME score.

## What is the role of Hindi swear words?

See top 10 words with the highest LIME scores, when predicting negative emotion.

| Token | Lang_ID | Swear Word?[2] |
|---|---|---|
| Fuck | eng | Yes |
| Chutiye | hin | Yes |
| Fakeionist | eng | No |
| Bsdk | hin | Yes |
| Sadly | eng | No |
| Bakwas | hin | No |
| Kutta | hin | Yes |
| Gaddar | hin | No |
| Shame | eng | No |
| Sala | hin | Yes |

Top 10 tokens with the highest LIME scores when predicting negative emotions, (anger, sadness, disgust and fear) for all models. They have been mapped to a canonical form and are in descending order of LIME score.

## What is the role of Hindi swear words?

See top 10 words with the highest LIME scores, when predicting negative emotion.

4/10 are Hindi swear words!

| Token | Lang_ID | Swear Word?[2] |
|---|---|---|
| Fuck | eng | Yes |
| Chutiye | hin | Yes |
| Fakeionist | eng | No |
| Bsdk | hin | Yes |
| Sadly | eng | No |
| Bakwas | hin | No |
| Kutta | hin | Yes |
| Gaddar | hin | No |
| Shame | eng | No |
| Sala | hin | Yes |

Top 10 tokens with the highest LIME scores when predicting negative emotions, (anger, sadness, disgust and fear) for all models. They have been mapped to a canonical form and are in descending order of LIME score.

# Do PLMs overgeneralise these associations?

# Do PLMs overgeneralise these associations?

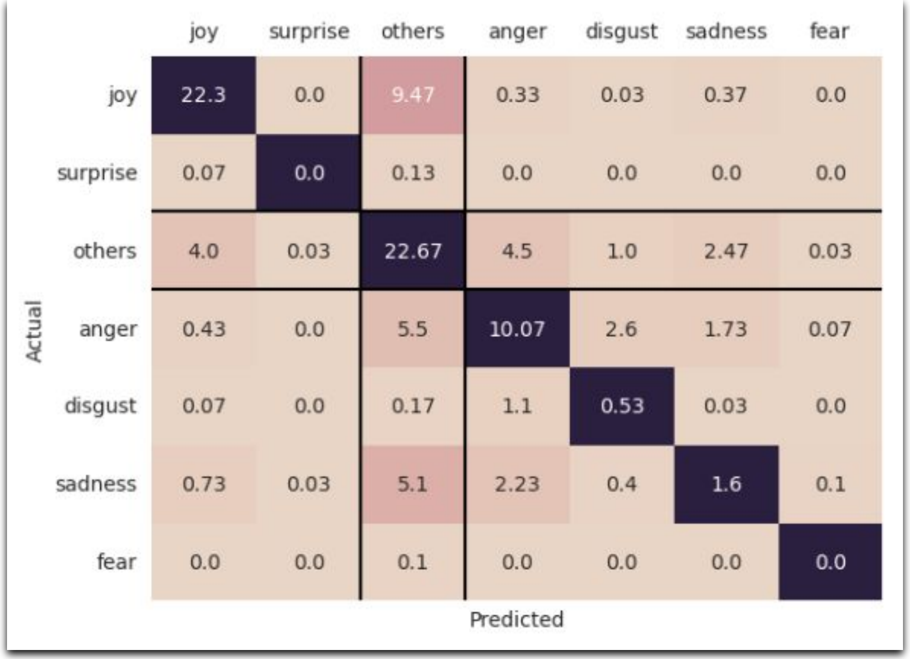
- Language models can adapt to heuristics that are valid for frequent cases and fail on the less frequent ones (McCoy et al, 2019).

# Do PLMs overgeneralise these associations?

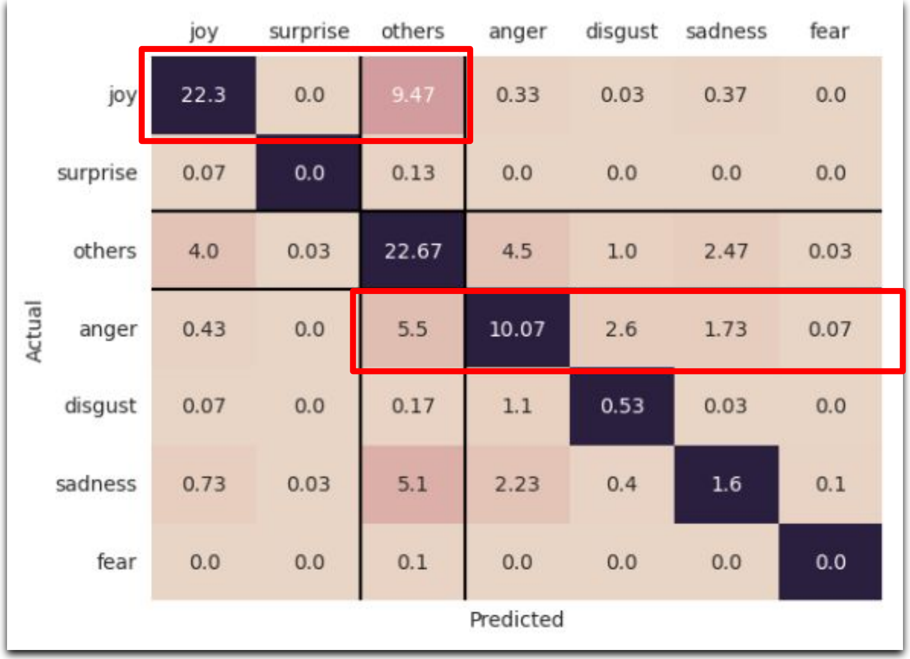
- Language models can adapt to heuristics that are valid for frequent cases and fail on the less frequent ones (McCoy et al, 2019).
- Will the sociolinguistics generalize to data-poor scenarios?

# Do PLMs overgeneralise these associations?

- Language models can adapt to heuristics that are valid for frequent cases and fail on the less frequent ones (McCoy et al, 2019).
- Will the sociolinguistics generalize to data-poor scenarios?
- We examine instances where the models have misclassified examples labelled as *joy* and *anger*.

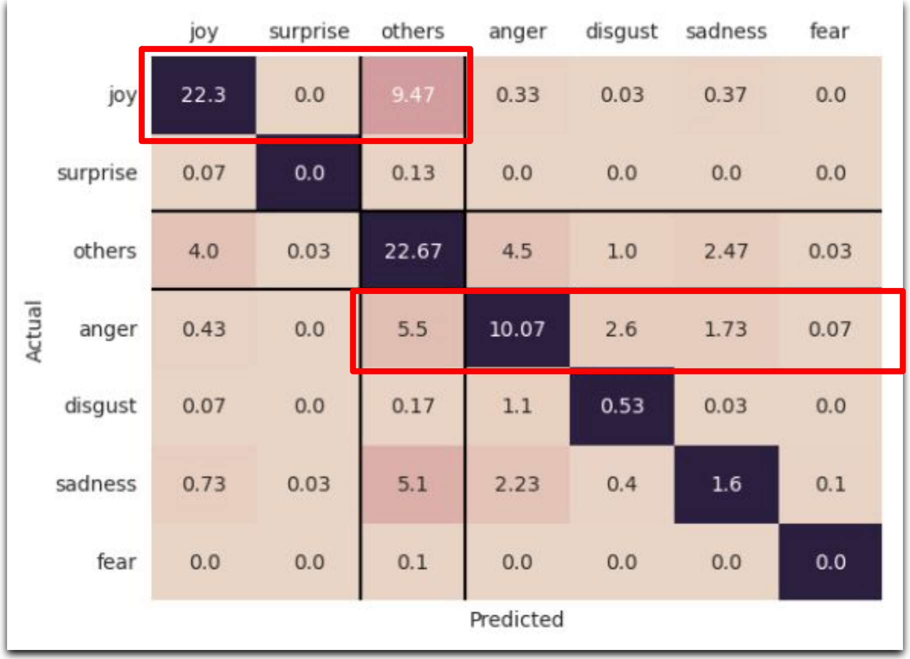



Confusion matrix containing the percentage of correctly and incorrectly classified examples for each label combination. The blue cells represent correct classifications, and the pink cells represent incorrect classifications.




Confusion matrix containing the percentage of correctly and incorrectly classified examples for each label combination. The blue cells represent correct classifications, and the pink cells represent incorrect classifications.

- Generally misclassified as another label of same emotional polarity.




Confusion matrix containing the percentage of correctly and incorrectly classified examples for each label combination. The blue cells represent correct classifications, and the pink cells represent incorrect classifications.

- Generally misclassified as another label of same emotional polarity.
- Models struggle with granular distinctions.


Confusion matrix containing the percentage of correctly and incorrectly classified examples for each label combination. The blue cells represent correct classifications, and the pink cells represent incorrect classifications.

- Generally misclassified as another label of same emotional polarity.
- Models struggle with granular distinctions.
- We also manually examine the few instances where this is not the case.

# Misclassifications

**Tweet:** @handle Very nice Sir yeh diya sateek jawab Pakistan ab bhi sudhar ja nahi to terey yaha sai jitn …

**Label:** Anger **Prediction:** Joy

An example labelled anger that was misclassified as joy owing to the English phrase (English - purple; Hindi - green; Other - orange) in the sentence having a positive connotation, even though the sentence itself conveys anger.

# Misclassifications

**Tweet:** @handle $\boxed{\text{Very nice Sir}}$ yeh diya sateek jawab Pakistan ab bhi sudhar ja nahi to terey yaha sai jitn …

**Label:** Anger **Prediction:** Joy

An example labelled anger that was misclassified as joy owing to the English phrase (English - purple; Hindi - green; Other - orange) in the sentence having a positive connotation, even though the sentence itself conveys anger.

# Misclassifications

- Distribution of English, Hindi, Other in misclassified examples.

| Joy | | | |
|---|---|---|---|
| Distribution of tokens in all examples | | | |
| | **All examples** | **Correct** | **Misclassified** |
| English | 0.40 | 0.44 | **0.32** |
| Hindi | 0.34 | 0.29 | **0.44** |
| Other | 0.26 | 0.27 | **0.24** |
| Distribution of tokens assigned a positive LIME score | | | |
| | **All examples** | **Correct** | **Misclassified** |
| English | 0.43 | 0.48 | **0.32** |
| Hindi | 0.32 | 0.28 | **0.42** |
| Other | 0.25 | 0.24 | **0.32** |
| **Anger** | | | |
| Distribution of tokens in all examples | | | |
| | **All examples** | **Correct** | **Misclassified** |
| English | 0.15 | 0.14 | **0.17** |
| Hindi | 0.63 | 0.65 | **0.61** |
| Other | 0.22 | 0.21 | **0.22** |
| Distribution of tokens assigned a positive LIME score | | | |
| | **All examples** | **Correct** | **Misclassified** |
| English | 0.15 | 0.13 | **0.18** |
| Hindi | 0.64 | 0.68 | **0.60** |
| Other | 0.21 | 0.19 | **0.23** |

# Misclassifications

- Distribution of English, Hindi, Other in misclassified examples.
- Misclassified *joy:* more **Hindi** tokens = more Hindi tokens have a higher LIME score.

| Joy | | | |
|---|---|---|---|
| Distribution of tokens in all examples | | | |
| | **All examples** | **Correct** | **Misclassified** |
| English | 0.40 | 0.44 | **0.32** |
| Hindi | 0.34 | 0.29 | **0.44** |
| Other | 0.26 | 0.27 | **0.24** |
| Distribution of tokens assigned a positive LIME score | | | |
| | **All examples** | **Correct** | **Misclassified** |
| English | 0.43 | 0.48 | **0.32** |
| Hindi | 0.32 | 0.28 | **0.42** |
| Other | 0.25 | 0.24 | **0.32** |
| **Anger** | | | |
| Distribution of tokens in all examples | | | |
| | **All examples** | **Correct** | **Misclassified** |
| English | 0.15 | 0.14 | **0.17** |
| Hindi | 0.63 | 0.65 | **0.61** |
| Other | 0.22 | 0.21 | **0.22** |
| Distribution of tokens assigned a positive LIME score | | | |
| | **All examples** | **Correct** | **Misclassified** |
| English | 0.15 | 0.13 | **0.18** |
| Hindi | 0.64 | 0.68 | **0.60** |
| Other | 0.21 | 0.19 | **0.23** |

# Misclassifications

- Distribution of English, Hindi, Other in misclassified examples.
- Misclassified *joy:* more **Hindi** tokens = more Hindi tokens have a higher LIME score.

| | All examples | Correct | Misclassified |
|---|---|---|---|
| **Joy** | | | |
| Distribution of tokens in all examples | | | |
| | **All examples** | **Correct** | **Misclassified** |
| English | 0.40 | 0.44 | **0.32** |
| Hindi | 0.34 | 0.29 | **0.44** |
| Other | 0.26 | 0.27 | **0.24** |
| Distribution of tokens assigned a positive LIME score | | | |
| | **All examples** | **Correct** | **Misclassified** |
| English | 0.43 | 0.48 | **0.32** |
| Hindi | 0.32 | 0.28 | **0.42** |
| Other | 0.25 | 0.24 | **0.32** |
| **Anger** | | | |
| Distribution of tokens in all examples | | | |
| | **All examples** | **Correct** | **Misclassified** |
| English | 0.15 | 0.14 | **0.17** |
| Hindi | 0.63 | 0.65 | **0.61** |
| Other | 0.22 | 0.21 | **0.22** |
| Distribution of tokens assigned a positive LIME score | | | |
| | **All examples** | **Correct** | **Misclassified** |
| English | 0.15 | 0.13 | **0.18** |
| Hindi | 0.64 | 0.68 | **0.60** |
| Other | 0.21 | 0.19 | **0.23** |

# Misclassifications

- Distribution of English, Hindi, Other in misclassified examples.
- Misclassified *joy:* more **Hindi** tokens = more Hindi tokens have a higher LIME score.
- Misclassified *anger:* more **English** tokens = more English tokens have a higher LIME score.

| **Joy** | | | |
|---|---|---|---|
| Distribution of tokens in all examples | | | |
| | **All examples** | **Correct** | **Misclassified** |
| English | 0.40 | 0.44 | **0.32** |
| Hindi | 0.34 | 0.29 | **0.44** |
| Other | 0.26 | 0.27 | **0.24** |
| Distribution of tokens assigned a positive LIME score | | | |
| | **All examples** | **Correct** | **Misclassified** |
| English | 0.43 | 0.48 | **0.32** |
| Hindi | 0.32 | 0.28 | **0.42** |
| Other | 0.25 | 0.24 | **0.32** |
| **Anger** | | | |
| Distribution of tokens in all examples | | | |
| | **All examples** | **Correct** | **Misclassified** |
| English | 0.15 | 0.14 | **0.17** |
| Hindi | 0.63 | 0.65 | **0.61** |
| Other | 0.22 | 0.21 | **0.22** |
| Distribution of tokens assigned a positive LIME score | | | |
| | **All examples** | **Correct** | **Misclassified** |
| English | 0.15 | 0.13 | **0.18** |
| Hindi | 0.64 | 0.68 | **0.60** |
| Other | 0.21 | 0.19 | **0.23** |

# Misclassifications

- Distribution of English, Hindi, Other in misclassified examples.
- Misclassified *joy:* more **Hindi** tokens = more Hindi tokens have a higher LIME score.
- Misclassified *anger:* more **English** tokens = more English tokens have a higher LIME score.

| | | | |
|---|---|---|---|
| **Joy** | | | |
| Distribution of tokens in all examples | | | |
| | **All examples** | **Correct** | **Misclassified** |
| English | 0.40 | 0.44 | **0.32** |
| Hindi | 0.34 | 0.29 | **0.44** |
| Other | 0.26 | 0.27 | **0.24** |
| Distribution of tokens assigned a positive LIME score | | | |
| | **All examples** | **Correct** | **Misclassified** |
| English | 0.43 | 0.48 | **0.32** |
| Hindi | 0.32 | 0.28 | **0.42** |
| Other | 0.25 | 0.24 | **0.32** |
| **Anger** | | | |
| Distribution of tokens in all examples | | | |
| | **All examples** | **Correct** | **Misclassified** |
| English | 0.15 | 0.14 | **0.17** |
| Hindi | 0.63 | 0.65 | **0.61** |
| Other | 0.22 | 0.21 | **0.22** |
| Distribution of tokens assigned a positive LIME score | | | |
| | **All examples** | **Correct** | **Misclassified** |
| English | 0.15 | 0.13 | **0.18** |
| Hindi | 0.64 | 0.68 | **0.60** |
| Other | 0.21 | 0.19 | **0.23** |

# To Conclude…

- PLMs learn sociolinguistic patterns established by literature when predicting emotion.

# To Conclude…

- PLMs learn sociolinguistic patterns established by literature when predicting emotion.
- Hindi = negative emotion; English = positive emotion.

# To Conclude…

- PLMs learn sociolinguistic patterns established by literature when predicting emotion.
- Hindi = negative emotion; English = positive emotion.
- PLMs can overgeneralise to infrequent examples.

# To Conclude…

- PLMs learn sociolinguistic patterns established by literature when predicting emotion.
- Hindi = negative emotion; English = positive emotion.
- PLMs can overgeneralise to infrequent examples.
- Motivation for deeper engagement between language model interpretability and sociolinguistics.

# To Conclude…

- PLMs learn sociolinguistic patterns established by literature when predicting emotion.
- Hindi = negative emotion; English = positive emotion.
- PLMs can overgeneralise to infrequent examples.
- Motivation for deeper engagement between language model interpretability and sociolinguistics.
- Future work: these understandings can be leveraged to make better systems designed for code-mixed languages.

# Selected Bibliography


- P. Agarwal, A. Sharma, J. Grover, M. Sikka, K. Rudra and M. Choudhury, "I may talk in English but gaali toh Hindi mein hi denge : A study of English-Hindi code-switching and swearing pattern on social networks," 2017 9th International Conference on Communication Systems and Networks (COMSNETS), Bengaluru, India, 2017, pp. 554-557, doi: 10.1109/COMSNETS.2017.7945452.
- A. Seza Doğruöz, Sunayana Sitaram, Barbara E. Bullock, and Almeida Jacqueline Toribio. 2021. A Survey of Code-switching: Linguistic and Social Perspectives for Language Technologies. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1654–1666, Online. Association for Computational Linguistics.
- S. Ghosh, A. Priyankar, A. Ekbal and P. Bhattacharyya. 2023. Multitasking of sentiment detection and emotion recognition in code-mixed Hinglish data. Knowledge-Based Systems, 260, 110182.
- Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.
- Marco Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, San Diego, California. Association for Computational Linguistics.
- Koustav Rudra, Shruti Rijhwani, Rafiya Begum, Kalika Bali, Monojit Choudhury, and Niloy Ganguly. 2016. Understanding Language Preference for Expression of Opinion and Sentiment: What do Hindi-English Speakers do on Twitter?. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1131–1141, Austin, Texas. Association for Computational Linguistics.




KU LEUVEN

Prediction: Joy

**Tweet:** @handle Wow dear I am proud of you kiya gali de ho aapne
**Lang_ID:** other eng eng eng eng eng eng eng hin hin hin hin hin
**Translation:** Wow, dear, I am proud of you. You have cursed so eloquently!

| | |
|---|---|
| **HingRoBERTa:** | @handle Wow dear I am proud of you kiya gali de ho aapne |
| **XLM-R:** | @handle Wow dear I am proud of you kiya gali de ho aapne |
| **IndicBERT:** | @handle Wow dear I am proud of you kiya gali de ho aapne |

All models are influenced by the English part of the sentence to predict *joy*